Speech-Driven Personalized Gesture Synthetics: Harnessing Automatic Fuzzy Feature Inference

Fan Zhang[1,2,3], Zhaohan Wang[4], Xin Lyu[4], Siyuan Zhao[1], Mengjian Li[3], Weidong Geng[5,3], Naye Ji[2], Hui Du[2],Fuxing Gao[2], Hao Wu[2], and Shunman Li[6]

[1]Faculty of Humanities and Arts, Macau University of Science and Technology, Macau, China, [2]The College of Media Engineering, Communication University of Zhejiang, China
[3]Research Center for Artificial Intelligence and Fine Arts, Zhejiang Lab, Zhejiang, China, [4]the School of Animation and Digital Arts Communication University of China, Beijing, China
[5]the College of Computer Science and Technology, Zhejiang University, Zhejiang, China, [6]Zhejiang Institute of Economics and Trade, Zhejiang, China

Introduction

- Speech-driven gesture generation is an emerging field within virtual human creation.
- A significant challenge lies in accurately determining and processing the multitude of input features (such as acoustic, semantic, emotional, personality, and even subtle unknown features).
- Traditional approaches, reliant on various explicit feature inputs and complex multimodal processing, constrain the expressiveness of resulting gestures and limit their applicability.
- **Persona-Gestor**, Our proposed model, generates personalized 3D fullbody gestures from raw speech alone, removing the need for manual style labels or feature extraction, as shown in Figure 1.

Contributions

1. We pioneering introduce the fuzzy feature inference strategy that enables driving a wider range of personalized gesture synthesis from speech audio alone, removing the need for style labels or extra inputs.
2. We combined AdaLN transformer architecture within the diffusion model to enhance the Modeling of the gesture-speech interplay.
3. Extensive subjective and objective evaluations reveal that our model is superior to the current state-of-the-art approaches.

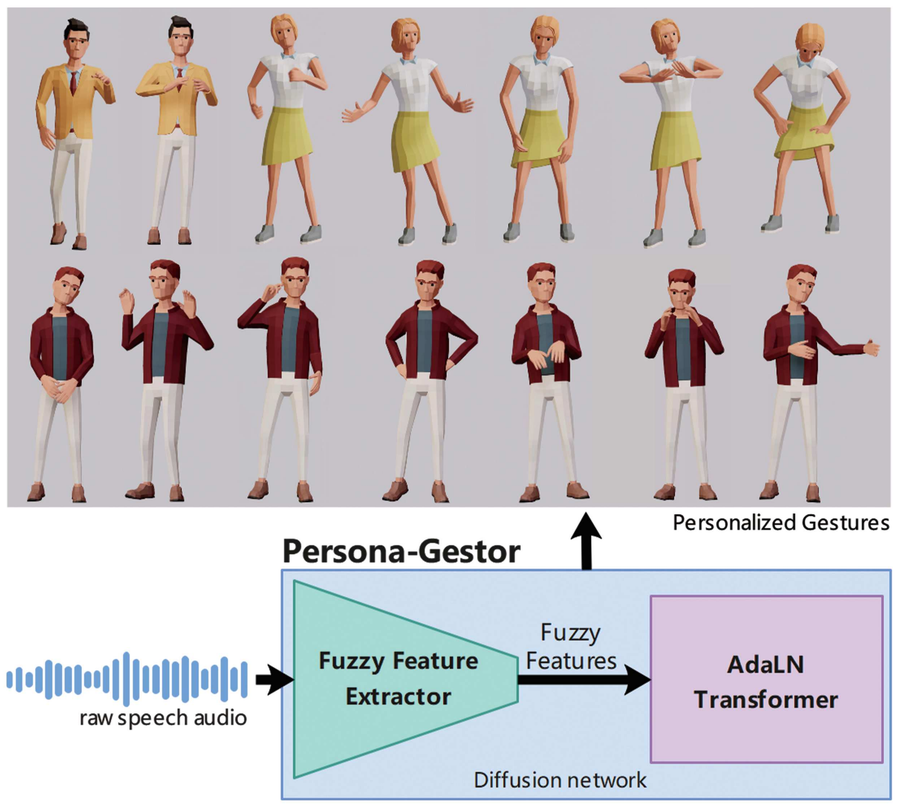

Figure 1. Each pose depicted is personalized gestures generated solely relying on raw speech audio. Persona-Gestor offers a versatile solution that bypasses complex multimodal processing and enhances user-friendliness.

System Overview

The model (Figure 2,3) combines a fuzzy feature extractor and a non-autoregressive Adaptive Layer Normalization (AdaLN) transformer diffusion architecture (DiTs-based).The fuzzy feature extractor harnesses a fuzzy inference strategy that automatically infers implicit, continuous fuzzy features. These fuzzy features, represented as a unified latent feature, are fed into the AdaLN transformer. The AdaLN transformer introduces a conditional mechanism that applies a uniform function across all tokens, thereby effectively modeling the correlation between the fuzzy features and the gesture sequence. This module ensures a high level of gesture-speech synchronization while preserving naturalness. we employ the diffusion model to train and infer various gestures.

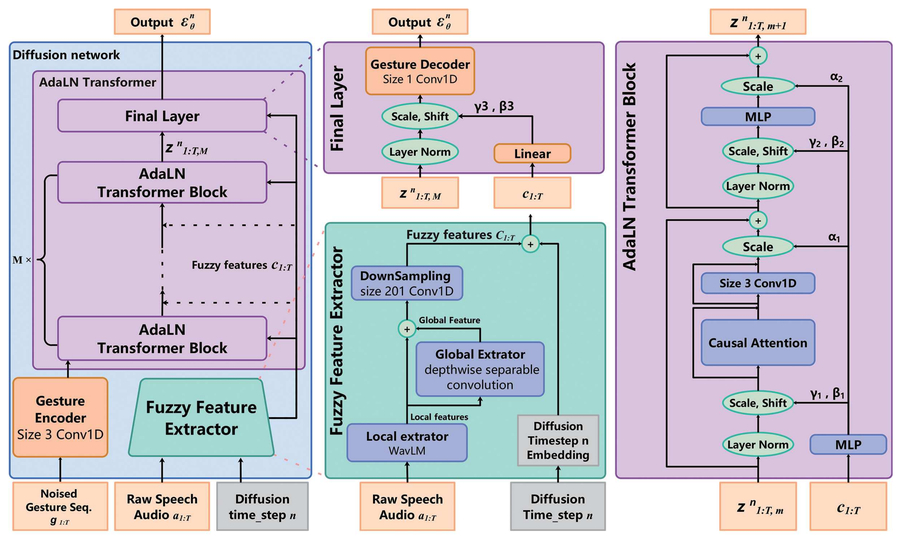

Figure 2. The architecture of Persona-Gestor.

Training & Inferencing with Denoising Diffusion Probabilistic Model

The model contains two processes: the diffusion process and the generation process.

At training time, the diffusion process gradually converts the original gesture data(g^0) to whitenoise(g^N) by optimizing a variational bound on the data likelihood.

At inference time, the generation process recovers the data by reversing this noising process through the Markov chain usingLangevin sampling.

Algorithm 1: Training for the Whole Sequence Gesture.

Input: data $g^0_{1:T} \sim p(g^0|a_{1:T})$ and $a_{1:T}$
repeat
  Initialize $n \sim \text{Uniform}(1,\ldots,N)$ and $\epsilon \sim \mathcal{N}(0,I)$
  Take the gradient step on

$$\nabla_\theta ||\epsilon - \epsilon_\theta \left(\sqrt{\overline{\alpha}_n} g^0_{1:T} + \sqrt{1-\overline{\alpha}_n}\epsilon, a_{1:T}, n\right)||^2$$

until *converged* ;

Algorithm 2: Sampling $g^0_{1:T}$ via Annealed Langevin Dynamics.

Input: noise $g^N_{1:T} \sim \mathcal{N}(0,I)$ and raw audio waveform $a_{1:T}$
for $n = N$ **to** 1 **do**
  if $n > 1$ **then**
    $z \sim \mathcal{N}(0,I)$
  else
    $z = 0$
  end if
  $g^{n-1}_{1:T} = \frac{1}{\sqrt{\alpha^n}}(g^n_{1:T} - \frac{\beta^n}{\sqrt{1-\overline{\alpha}^n}}\epsilon_\theta(g^n_{1:T}, a_{1:T}, n)) + \sqrt{\sigma_\theta} z$
end for
Return $g^0_{1:T}$

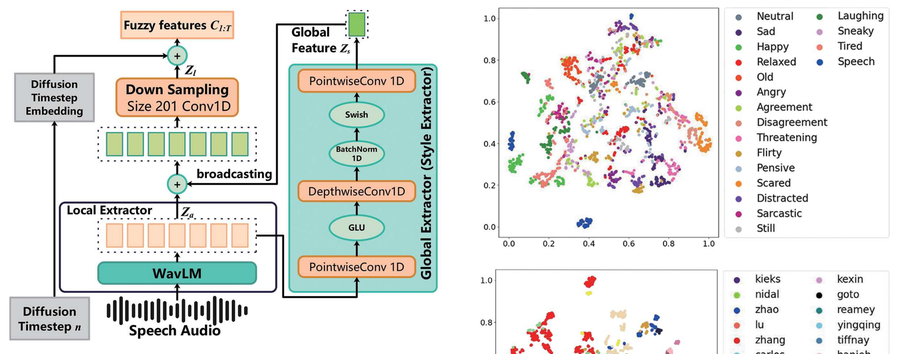

Figure 3. An overview of the fuzzy inference condition extractor.

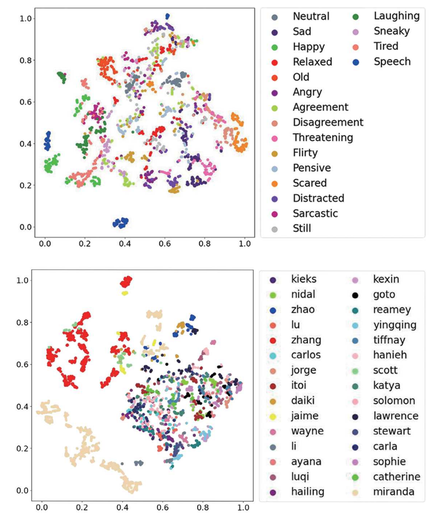

Figure 5. The T-SNE clustering visualization displays a variety of gesturesdistinguished by color-coded styles, revealing distinct regions for gestures tiedto specific emotions or speakers, albeit with some boundary overlaps. Thishighlights our approach's capability to produce distinct style gestures through afuzzy feature inference strategy, relying solely on speech audio.

Visulization Results

Our system excels in creating personalized gestures that align with the speech context, leveraging the fuzzy inference strategyto autonomously derive fuzzy features directly from speech audio. Furthermore, it showcases remarkable generalization and robustness by utilizing in-the-wild speech, as shown in Figure 4 and 5.

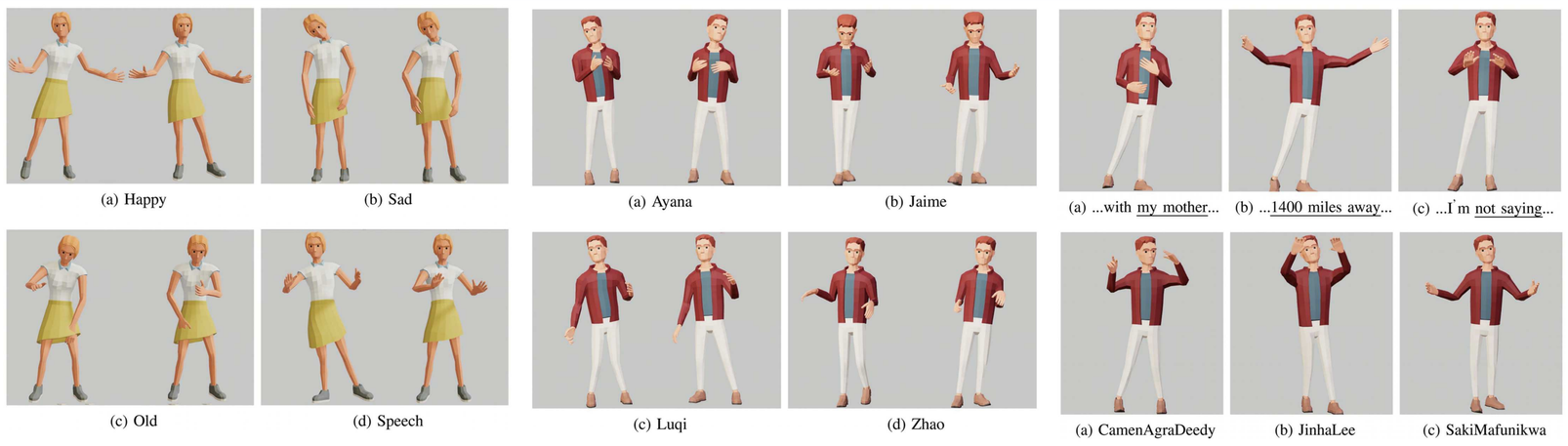

Figure 4. The visulization results of gestures corresponding to different emotions(left), personalities(middle),and semantic(right).

Subjective & Objective Evaluation

- **Dataset:** We conducted experimental evaluations using the Trinity, ZEGGS, and BEATS datasets.
- **Subjective Evaluation Metrics:** Human-likeness, Appropriateness, and Style appropriateness.
- **Objective Evaluation Metrics:** FrÃchet Gesture Distance (FGD) and BeatAlign.
- Extensive subjective and objective evaluations confirm our model's superior performance to the current state-of-the-art approaches, shown in Table 1.

Table 1. The subjective & Objective Score.

Dataset	Methods: Model	With Fingers	Subject Evaluation Metric: Human↑ likeness	appropriateness↑	Style↑ appropriateness	Objective Evaluation Metric: FGD↓ on feature space	FGD↓ on raw data space	BeatAlign↑
Trinity	GT	*Y*	0.51±1.73	0.66±1.24	/	/	/	/
	LDA [4]	*N*	-0.22±0.98	-0.39±1.08	-0.18±1.10	349.53	6008.37	0.68
	Taming [15]	*N*	-0.48±0.96	-0.47±1.07	-0.23±1.05	3970.14	52196.87	0.68
	(Proposed)PG	N	**0.12±1.09**	**0.19±1.12**	**0.20±1.06**	**289.42**	**5080.57**	**0.69**
	(Ours)PGNSE	N	0.08±1.06	0.13±1.18	0.15±1.03	8002.65	70617.28	0.68
	(Ours)PGCA	N	-0.17±1.02	-0.01±1.11	-0.03±1.02	2566.65	26124.39	0.68
	(Ours)PGCF	N	-0.08±1.06	-0.13±1.05	0±1.02	12540.17	156895.12	0.68
	(Ours)PGICC	N	0.07±1.01	0.06±1.12	0.08±1.01	125921.53	1940124.44	0.67
ZEGGS	GT	Y	0.95±1.13	1.19±1.03	/	/	/	/
	LDA [4]	N	-0.73±1.12	0.02±1.31	0.53±1.41	124.55	50996.33	0.66
	DSG [2]	Y	-0.47±1.14	-0.71±1.11	-0.61±1.08	66.77	33297.50	0.63
	Taming [15]	Y	-1.08±1.01	-0.91±1.09	-0.98±0.97	1419.76	293245.12	0.67
	ZeroEGGS [7]	Y	0.38±1.11	0.29±1.35	0.49±1.31	37.19	26666.85	0.66
	(Proposed)PG	Y	**0.42±1.17**	**0.48±1.29**	**0.76±1.34**	**28.13**	**26193.92**	**0.68**
	(Ours)PGNSE	Y	0.33±1.15	0.35±1.29	0.59±1.38	125.40	49081.55	0.66
	(Ours)PGOnehot	Y	0.25±1.19	0.33±1.31	0.51±1.28	122.56	50259.95	0.63
	(Ours)PGCA	Y	-0.36±1.22	-0.62±1.14	-0.66±1.09	807.12	156686.01	0.67
	(Ours)PGCF	Y	0.27±1.20	-0.03±1.28	-0.23±1.26	97.57	39256.55	0.67
	(Ours)PGICC	Y	0.04±1.20	-0.39±1.23	-0.40±1.21	407.89	89893.96	0.67
BEAT	GT	Y	0.65±1.16	0.96±1.04	/	/	/	/
	LDA [4]	*N*	-1.65±0.73	-1.59±0.74	-1.35±1.05	276.25	3584.95	0.66
	DSG+ [3]	Y	-0.28±1.17	-0.49±1.15	-0.40±1.24	23811.46	2384465.64	0.43
	GDC [16]	*N*	0.54±1.12	0.47±1.25	0.30±1.27	432.15	93215.56	**0.69**
	Taming [15]	Y	-0.42±1.14	-0.52±1.14	-0.32±1.24	1251.56	46828.23	0.66
	(Proposed)PG	Y	**0.56±1.14**	**0.63±1.10**	**0.66±1.16**	**264.06**	**3471.26**	0.68
	(Ours)PGNSE	Y	0.09±1.16	0.27±1.23	0.46±1.31	1514.94	51077.98	0.66
	(Ours)PGOnehot	Y	-0.01±1.16	0.18±1.31	0.32±1.36	1863.69	63872.78	0.63
	(Ours)PGCA	Y	0.35±1.09	0.17±1.26	0.28±1.33	703.83	18990.56	0.66
	(Ours)PGCF	Y	0.14±1.01	0.15±1.22	0.30±1.31	1160.63	48899.63	0.66
	(Ours)PGICC	Y	0.02±1.10	-0.24±1.17	-0.25±1.25	2057.31	78754.92	0.66

This work has been published in IEEE Transactions on Visualization and Computer Graphics(TVCG), Vol. 30, No. 10, October 2024.

Supplementary videos and code can be accessed at https://zf223669.github.io/Diffmotion-v2-website/

Contact
- E-mail: FanZhang@cuz.edu.cn
- ORCiD: 000-0002-9534-1777